Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 318 – 327

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

A privacy preserving technique to prevent sensitive behavior exposure in semantic location-based service

Yuna Oh*, Kangsoo Jung, Seog Park

Sogang University, Department of Computer Engineering, Seoul, Republic of Korea

Abstract

The increasing number of mobile device users indicates the expansion of personalized location-based services (LBS). Despite their proliferation, the risk of violating users' privacy by exposing user's location information remains. For this reason, many studies have researched to prevent privacy violation in LBS. However, previous researches only focused on protecting users' location information without considering semantic location privacy violation through contextual information. In this paper, we explain the process of inferring a user's behavior using semantic information which includes spatial and temporal information. We also suggest a privacy preserving technique to prevent exposure of sensitive behavior in semantic LBS. We implement an android application to validate the proposed technique. In accordance with the experimental results, the proposed b-diversity technique is validated to prevent exposure of sensitive behavior and also minimizing data utilization degradation.

Keywords: semantic location;privacy;decision tree;b-diversity

1. Introduction

The growing number of mobile devices users and the development of wireless network have given rise to various personalized services. Location-based services (LBS) have gained popularity by providing personalization services such as navigation, online social networks and point of interest (POI) search. However, LBS have risk to expose a

* Corresponding author.
E-mail address: azure84@sogang.ac.kr

doi:10.1016/j.procs.2014.08.112

user's location history and contextual information to malicious adversary. For instance, user's political beliefs can be inferred by specific information such as duration of stay in a specific building occupied by political parties. Therefore, in order to provide safe and reliable location-based services, user's privacy should be guaranteed.

Numerous researches have been studied to solve the privacy issues arising from the use of LBS. Existing research to protect the location privacy can be classified into policy-based technique and inference prevention techniques. The policy-based technique is used to reflect the user agreement and privacy setting to decide level of information disclosure. Inference prevention technique is applied to prevent inference of sensitive information that the user does not want the public to expose. The inference prevention technique can be further categorized into three types, namely, identity privacy, location privacy, and semantic location privacy.

Identity privacy protection prevents the identification of a user who uses LBS and location privacy protection is the research to prevent to expose the exact position of the LBS user. Semantic location privacy protection [9-11] prevents the exposure of a user's contextual information associated with location information. Semantic location privacy protection is different from identity privacy and location privacy because the former considers location or identification information and contextual information, such as user's behavioral information. However, research on semantic location privacy technique is not yet studied sufficiently.

In this paper, we focus on a privacy protection technique to prevent the exposure of the sensitive behavioral information in semantic LBS. This technique uses the spatial and contextual information of the user to provide a personalized service. We propose a b-diversity method to reduce the probability of sensitive behavior inference to 1/b by maintaining b or more candidate behavior that can be inferred.

The remainder of this paper is organized as follows: Section 2 reviews existing LBS privacy and semantic location privacy techniques. Section 3 introduces the proposed b-diversity method and algorithm. Section 4 explains our experiment to validate the proposed technique. The last section summarizes the contributions of this paper and our future work.

2. Related works

2.1. Location privacy preserving techniques

Privacy preserving techniques in LBS are classified into location anonymization which hides a user's exact location information, and identity anonymization which prevents user identification. Location k-anonymity takes k-1 other user's locations to reduce the possibility of pinpointing a user's exact location. This method is the most popular technique for location anonymization. Location anonymization techniques are further classified into spatial cloaking[1,2,3], space transformation[4,5], and fake location[6,7].

Spatial cloaking has been widely used to tackle privacy issues in LBS. The basic idea of spatial cloaking techniques is to blur or generalize a user's exact location into a cloaked area to satisfy location k-anonymity. Spatial cloaking is a simple technique; however it deteriorates the quality of service and require additional computation cost. Space transformation is a method that transforms a user's location into another space to hide his/her location, while maintaining the spatial relationship. This method is more efficient than spatial cloaking; however the calculations involved are complex, and are unable to support various query types, such as range query. Fake location technique sends dummy, false location information, to the LBS provider to conceal a user's location. The fake location does not deteriorate the quality of service, but disclosing dummy information is possible. Moreover, the creation and management of dummies is expensive.

Identity anonymization[8] uses pseudonyms to prevent user identification by hiding information, such as name and Social Security Number (SSN). However, if a certain pseudonym is used for a long time, it can eventually be inferred. A previous work[8] has introduced the concept of mix zones. A mix zone is an area that assigns new pseudonyms without exposing the relationship between the user's old and new pseudonyms.2.2

2.2. Semantic location privacy

In an LBS service environment, semantic location is defined to generate new semantic information comprising user preferences and contextual information on specific space. LBS using semantic location can provide various

personalized services that reflect the tastes and preferences of individual users. However, the increase in personalization also increases privacy violation risks.

To ensure privacy protection of semantic location, a previous work[9] has proposed cloaking technique that satisfies a user's privacy preference. In this study, the probability of a user's location and sensitivity threshold is considered when deciding one the cloaking area. Yigitoglu[10] proposed a cloaking technique that consider the road and rail network. This technique is distinct from other existing techniques have no limitation for object movement; it also considers contextual information such as location type and minimum time to reach specific location. It creates cloaking region by calculating the sensitivity and probability of specific locations. Lee[11] regards residence time as a characteristic of the semantic location and presents a cloaking area generation technique to prevent the inference of the user's exact location. This method assumes that the purpose of one's to a location can be inferred based on the residence time. For instance, visiting a restaurant for a meal usually takes 1 to 2 hours, and staying in school to attend classes takes 5 to 6 hours. Lee[11] proposed a technique for generating a clocking area that can prevent to infer the purpose of visit the place through the difference of residence time.

Existing LBS privacy preserving techniques aim to prevent identification and location information leakage. However, semantic location privacy focuses on preventing user's semantic information exposure such as behavioral information of purpose of visit. Consequently, existing LBS privacy protection technique cannot be applied to semantic location privacy.

In this paper, we propose a semantic location privacy preserving technique that generates cloaking area using user's context including spatial and temporal information. This technique is used to prevent inferences with respect to user's sensitive behavior that user want to keep private.

3. Privacy preserving technique for semantic location

3.1. Motivation

In general, each location has a unique behavior based on location's type. For instance, specific behavior that related with health care occurs in hospitals. In government offices, public service activities are performed. Each unique behavior has a pattern, such as residence time, frequency, in/out time, and unique behaviors, which can be distinguished using these patterns. For instance, if a certain user's residence time pattern is usually 6 hours and his in/out time pattern is 09:00 to 18:00, we can infer that the user is someone who performs medical related work. Hence, the unique behavior is determined by the location type, and such behaviors can be inferred based on their own pattern. Table 1 presents an example of unique behaviors in the medical area, a field that is considered to be sensitive to issues. This table is obtained from a survey in the medical field.

Table 1 An example of unique behaviors in the hospital

Unique behavior	In/out time	Residence time	Frequency
Visit for medical treatment by patients	09:00-18:00	2-4 hour	Aperiodic
Visits as doctor	00:00-24:00	Almost 10 hour	Periodic
Visits as nurse	00:00-24:00	Almost 8 hour	Periodic
Patients admitted to hospital	00:00-24:00	24 hours	Aperiodic
Visit for medical receptionist	09:00-18:00	1 hours	Aperiodic

Personalized services based on the semantic location of the user can allow managers to infer the users' unique behavior. It is possible to provide a convenient personalized service, but at the same time, this inference occur serious privacy violation. Therefore, semantic location privacy preserving technique which can control the degree of a user's behavior exposure, is required.

3.2. Basic scheme

Lee[11] has presented the possibility of inferring a user's current location using residence time pattern; it also proposes a cloaking algorithm to prevent the exposure of the user's location to be inferred. This method attempts to prevent the expose of the user's location. However, inference of a unique behavior in a specific location might be more dangerous than disclosure of location. In this paper, we define various unique behaviors in a semantic location. Then, the sensitivity of unique behavior is set by a user in specific location to meet each user's different privacy requirements. We use a decision tree algorithm to infer unique behavior using collected information and apply the proposed b-diversity method to prevent sensitive behavior. An overview of the proposed technique is shown as follows.

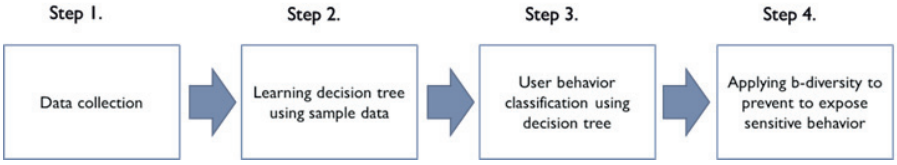

Fig.1. Overview of the proposed technique

We assume that the generation of a decision tree is performed by a trusted third party. The resulting decision tree is broadcasted to each user's mobile device. The decision tree can be generated by distributed method using SMC without a trusted third party; however we provide no detailed explanation here because it is beyond the scope of this study. The user infers her behavior before reporting her information to the LBS provider. If the classification result is classified as sensitive behavior, the user adds noise or creates cloaking area to hide the sensitive behavior from the LBS provider.

3.3. Decision tree generation for behavior classification

The decision tree, which represents as a tree-like structure, analyzes data to generate a rule for classification. Using the decision tree algorithm has several advantages. First, this method is easy to understand because it provides clear reason for the classification. Second, when data have too many attributes, the decision tree can automatically and efficiently exclude the attributes that do not affect the results. In addition, this method can save time and effort because it can process data without transforming data value.

Several decision tree algorithms depend on the utility function such as ID3, CART, CHAID, C5.0. In this paper, we have chosen the ID3 algorithm, which uses information entropy to calculate information gain, because this is typically used to generate the decision tree. Clearly, other decision tree algorithms can be used.

We implemented an android application to collect data including spatial and temporal information. We collected contextual information from 10 users for 10 days. The collected data samples are listed in table 2.

Table 2 An example of collected data

Type of location	Residence Time	Time slot	Day of week	Latitude	Longitude	Time
Education	1hour	Afternoon	Thur.	37.55	126.94	17:42
Restaurant	1 hour	Evening	Fri.	37.49	127.13	20:29
Café	2 hour	Evening	Sat.	37.48	127.03	20:52
Office	2 hour	Morning	Sun.	37.494	127.12	10:35

To generate the decision tree, each user manually attaches a label as behavior information to their collected data via application. This method generates additional cognitive cost; hence frequency and granularity of input label should be adjusted for service environment. When labeling the behavior, the user also inputs sensitivity of behavior to their data. The sensitivity scale is 1 to 10. In this scale, 1 means the lowest sensitivity and 10 represent the highest sensitivity value. Sensitivity value of behavior is used to calculate the average sensitivity of each behavior. We

predefine a set of possible behaviors for labeling. If the user wants to input other behaviors that are not predefined, they can input the label manually. We generate decision trees based on the collected data, as shown in is Fig.2

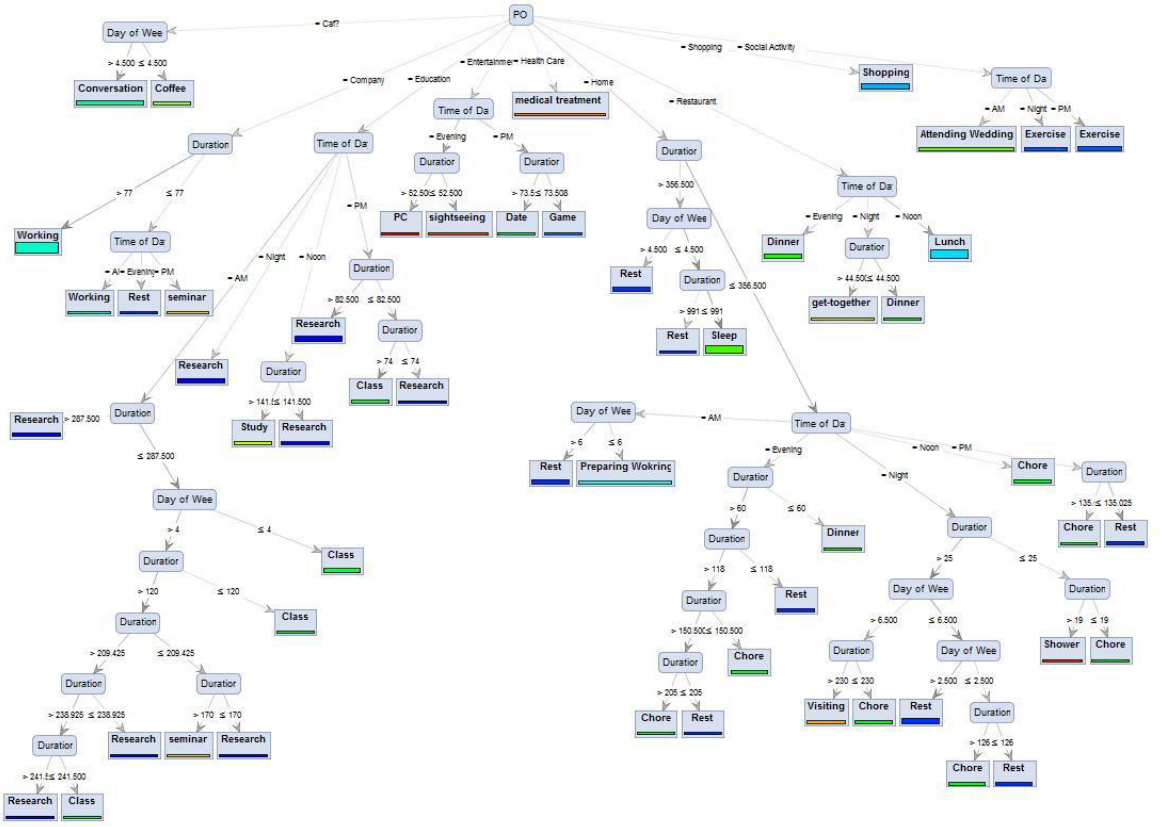

Fig. 2. Decision tree based on collected data

3.4. Classification using decision tree

Some behaviors classified using the decision tree are considered user sensitive. The proposed technique focuses on preventing the exposure of sensitive behavior in specific locations. For instance, we assume that Alice knows that her behavior is inferred by her data using a decision tree [Table 3]. Alice does not care about the exposure of location information such as school for learning or pub for social activity. However, she wants to prevent exposure of her sensitive behavior such as hospital visiont to take medical treatment. The existing cloaking method simply attempts to protect her location privacy. Hence, this method is inappropriate to meet her requirement. Additionally, the cloaking method can lead to an extreme deterioration of data utilization.

Table 3 Classified behavior of Alice

Type of location	Residence Time	Time slot	Day of week	Time	Behavior
Hospital	2 hours	Morning	Thur.	10:35	Medical treatment
Hospital	6 hours	Afternoon	Fri.	15:29	Medical treatment
School	5 hours	Morning	Sat.	10:52	Learning
Pub	2 hours	Night	Sun.	20:35	Social activity

3.5. Classification using decision tree

Behavior sensitivity

Behavior sensitivity is used to estimate the degree of behavior's privacy. As we have explained in section 3.3, the sensitivity of each behavior is calculated based on user input. Every behavior has a default sensitivity value, which is the average of the entire range of sensitivity values displayed by the user. In addition, users can change default sensitivity value to reflect his privacy preference and even set the sensitivity threshold. If the inferred behavior's sensitivity value exceeds the sensitivity threshold, the behavior is defined as sensitive behavior, and b-diversity is applied to prevent the exposure of such sensitive behavior.

The b-diversity algorithm

If the sensitive behavior, which exceeds the sensitivity threshold, is inferred by the decision tree, then the proposed b-diversity method adds noise to user's spatial and temporal data to infer other b-1 behaviors using decision tree. The value of the average sensitivity of the inferred behavior should be less than that of the sensitivity threshold. Adding noise to infer other b-1 behaviors results in the reduction of the probability of exact behavior inferred less than 1/b. In this paper, we named proposed technique as the b-diversity.

Definition 1. b(behavior)-diversity

The proposed b-diversity is used to protect a user's semantic location privacy by adding noise or generating cloaking area to infer b or more behaviors, which reduces the probability that user's behavior can be inferred to less than 1/b.

Generating cloaking area significantly decreases data utility in LBS. However, existing location privacy protection techniques, including semantic location privacy, simply use the cloaking method to preserve privacy. The proposed b-diversity technique adds noise to infer other b-1 behaviors before generating the cloaking area to prevent exposure of sensitive behavior. This proposed technique can reduce data utilization by constraining cloaking area generation. The b-diversity algorithms is presented below.

1. The user checks whether the inferred behavior exceeds the sensitivity threshold before sending data to the LBS provider. If the inferred behavior is non-sensitive, the user reports data to the LBS provider. If the behavior is sensitive, the user applies b-diversity algorithm to prevent exposure of such sensitive behavior.
2. The b-diversity algorithm initially attempts to selects the inferred behavior's sibling behavior in decision tree, in accordance with an ascending order of sensitivity value, in order to decrease average sensitive of behaviors.
3. The proposed algorithm adds noise according to the decision rule to generate more data than those inferred from the b-1 inferred behavior.
4. If no sibling behavior is observed, and the average sensitivity value is less than the threshold, we attempt to find another behavior to meet the threshold constraint in the parent node.
5. If the decision rule's attribute is the location attribute, then no candidate behavior is added in the current location. In this case, we generate the cloaking area to find the candidate behavior in another location.
6. Steps 2 and 5 are repeated until more b-1 behaviors are found to satisfy the threshold constraint.

In the b-diversity algorithm, the value of b and the sensitivity threshold affect the data utility and privacy protection. As their values become increasingly larger and larger, the privacy protection degree increases. However, this degree deteriorates data utility by adding noise and generating the cloaking area. Hence, the values of b and sensitive threshold should be determined by considering a trade-off between data utility and degree of privacy protection degree.

Algorithm 1. b-diversity

```
Input:  b          //value of b-diversity, inferred behavior
Output:  CB[i]         //Candidate behavior
Candidate behavior  CB[i]; Decision tree's node  N_i ;
Sensitivity value of behavior  S_j; Threshold of sensitivity  T;
Current inferred behavior  B_i;

While(number of  CB[i]  !=b){
    for(j=0 to number of behavior in N_i){
        select the behavior S_j in B_i 's sibling node
        if(sensitivity of S_i>=T)
            continue;
        else if(sensitivity of S_i<T)
            CB[i]=S_i
         i++;
    }
    if(number of  CB[i]  !=b&&no candidate S_j in N_i)
        N_i=B_i 's parent node
}
Return  CB[i];
```

Fig. 3. The proposed b-diversity algorithm

4. Experiment

4.1. Experimental purpose

In this chapter, we validate whether the proposed technique can guarantee sufficient semantic location privacy protection and minimize data utilization degradation. We compared the proposed technique to [11], in terms of privacy protection and data utilization. We carried out three experimental evaluations. (1) cloaking area size by increasing sensitivity threshold, (2) cloaking area size by increasing number of sensitive behavior, and (3) cloaking area size by number of b. Cloaking area is significantly related with data utility and privacy protection. Hence, we compared the cloaking area size between existing semantic location privacy protection technique and proposed technique.

4.2. Experimental purpose

We performed the experiment using Intel® Core™ i5 3.1GHz CPU, 3GB RAM, and Windows 7 Home Premium K SP1. We implemented the Android application to collect the users' context information, including spatial and temporal information. We released our application to 10 people, and collected their information for 10 days automatically via released application. Fig. 3 shows an application interface and collected location information of users. We generated the decision tree using ID3 algorithm. We evaluate the accuracy of the decision tree by comparing a user's real behavior and classified behavior using the decision tree. Upon evaluation, the average accuracy of the generated decision tree is found to be 69.55%.

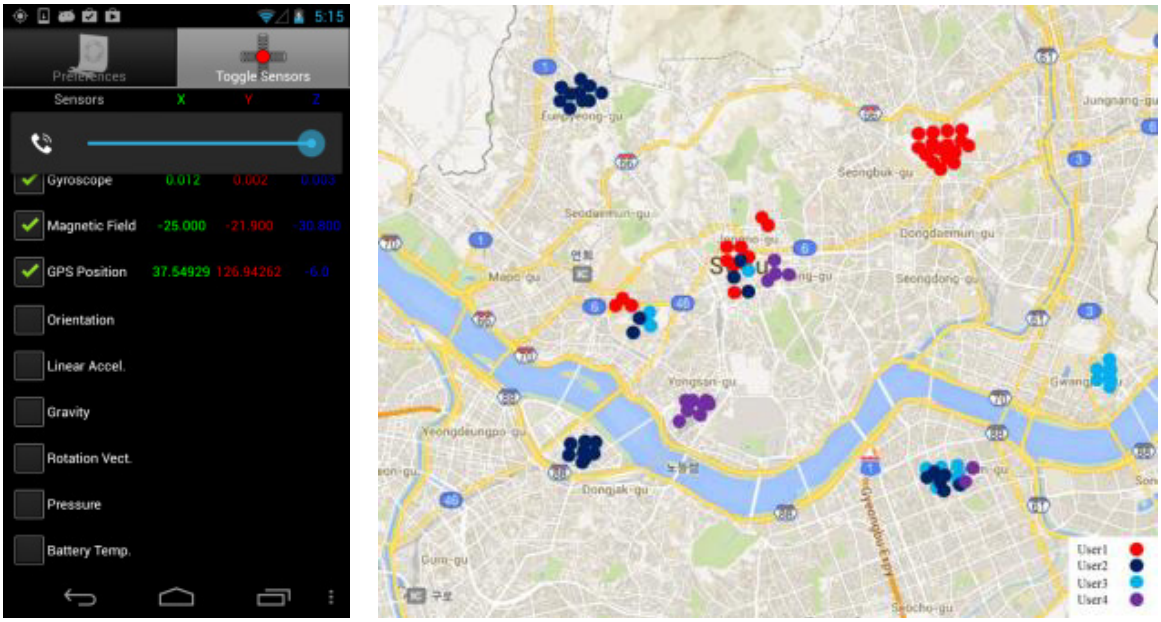

Fig.4. Collected user data and Android application for data collection

4.3. Experimental purpose

To validate the data utility of the proposed technique, we compared our technique to Lee[11] in terms of cloaking area size. Then, we evaluated the cloaking area size by increasing sensitivity threshold. The experiment result is shown in the left side of Fig.5, in which the X axis indicates sensitivity value threshold, and the Y axis represents the ratio of cloaking area size. Experiment result shows that the proposed technique's cloaking area size is 52.78% less than that in Lee[11]. In Lee[11], the authors generated cloaking area for the protection of the location privacy. However, the proposed technique adds noise to satisfy b-diversity. If b-diversity cannot be satisfied by adding noise, then the cloaking area is generated. Therefore, the proposed technique can prevent the increase of cloaking area size while minimizing data utility deterioration for privacy protection.

4.4. Experimental purpose

We evaluate the cloaking area size by increasing the number of sensitive behavior. We compare the proposed technique to Lee[11]. The experiment result is shown in the right side of Fig.5, in which X axis indicates the number of sensitive behavior/number of whole behavior, and the Y axis represents the ratio of cloaking area size. When the sensitive behavior ratio is 90%, the proposed technique's cloaking area size is 48% less than Lee[11] because the proposed technique considers b-diversity before generating the cloaking area. In LBS, location is the most significant factor for data utility. Hence the proposed technique prevents data utility degradation when the user needs high standard privacy protection.

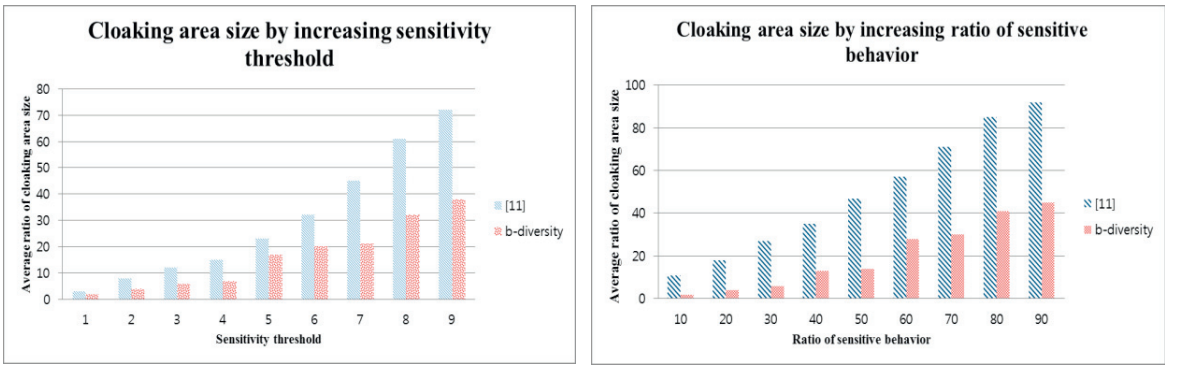

Fig. 5. Cloaking area size by increasing sensitivity threshold and ratio of sensitive behavior

4.5. Cloaking area size by increasing the number of b

We evaluated the cloaking area size by increasing the number of b, and the experiment result is shown in Fig. 6. We find that when b-diversity is 5, cloaking area size increases rapidly. On average, the number of candidate behavior is 4. If b is more than 4, it is more difficult to add noise in order to satisfy b-diversity. Consequently, the proposed technique generates cloaking area to meet the b-diversity requirement. We should select the b value depending on the average number of candidate behavior, because cloaking area generation significantly degrades data utilization.

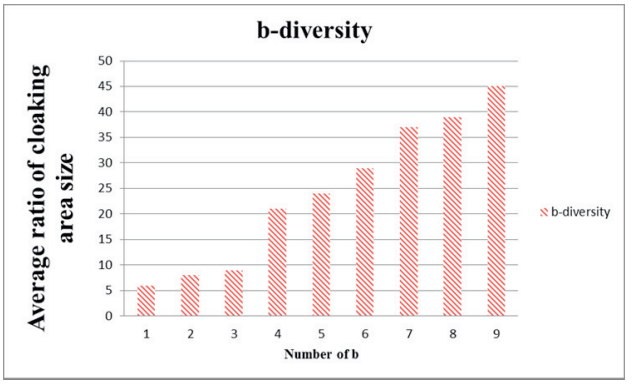

Fig. 6. Cloaking area size by increasing number of b

5. Conclusion

LBS have become more popular because of the increasing number of mobile device user. However, privacy violation using location information remains as the crucial weakness of LBS proliferation. Semantic location information considers a user's preference or context information, which is useful in providing personalized LBS. However, semantic location privacy protection remains insufficiently studied despite the risks involved in exposing user's location or identification information. Our contributions in this paper are as follows. First, we consider various spatial and temporal factors of semantic location information to preserve a user's privacy violation. Second, we propose the b-diversity technique to prevent the exposure of sensitive user behavior while minimizing data utility deterioration. To apply the proposed technique, we need additional cost and trusted third party server to generate and maintain the decision tree. However, such cost is reasonable when considering the need to preserve user privacy in LBS. In our future work, we will extend proposed technique to consider relationship among companions and location trace of each user. By these considerations, we can cover more various semantic location information that we should protect.

Acknowledgements

This research was supported by basic Science research program through the national research foundation of korea(NRF) funded by the ministry of education(2013R1A1A2013172)

References

1. Mokbel MF, Chow CY, Aref WG, The new casper: query processing for location services without compromising privacy, Proceedings of the 32nd international conference on Very large data bases, 2006. pp.763–74.
2. Chow CY, Mokbel M, Liu X, Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments, Journal of Geoinformatica, vol. 15, No.2, 2011.pp. 351-380.
3. Tanzima Hashem, Lars Kulik, Don't trust anyone: Privacy protection for location-based services, Pervasive and Mobile Computing, Vol.7, No.1, 2011.pp. 44–59.
4. Khoshgozaran A., Shahabim C, Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy, Proceedings of the 10th international conference on Advances in spatial and temporal databases, 2007.pp.239-257.
5. Pingley A. et al., CAP: A context-aware privacy protection system for location-based services, Proceedings of the 29th IEEE International Conference on Distributed Computing Systems, 2009.pp. 49–57.
6. Kido H, Yanagisawa Y, Satoh T, An anonymous communication technique using dummies for location-based services, Proceedings of the 1st International Conference on Pervasive Services, 2005. pp. 88-97.
7. Santos F, Humbert M, Shokri R, Hubaux JP, Collaborative location privacy with rational users, , Proceedings of the 2nd Conference on Decision and Game Theory for Security, 2011.pp.163-181.
8. Beresford A. Stajano F., Location privacy in pervasive computing," IEEE Pervasive Computing, Vol. 2, No.1, 2003.pp. 46–55.
9. Damiani ML et al. "Fine-Grained Cloaking of Sensitive Positions in Location- Sharing Applications", IEEE Pervasive Computing, Vol. 10, No.4, 2011.pp. 64–72.
10. Yigitoglu et al. "Privacy-preserving sharing of sensitive semantic locations under road-network constraints." Proceedings of the IEEE 13th International Conference on Mobile Data Management (MDM), 2012. pp.186-195.
11. Byoungyoung L. et al. "Protecting location privacy using location semantics" Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011. pp.1289-1297.